eip ESCUELA INTERNACIONAL POSGRADOS

Máster Avanzado de Programación en Python para Hacking, BigData y Machine Learning

Programación Python para BigData

LECCIÓN 07

Apache Spark con PySpark [1/2]

ÍNDICE

INTRODUCCIÓN

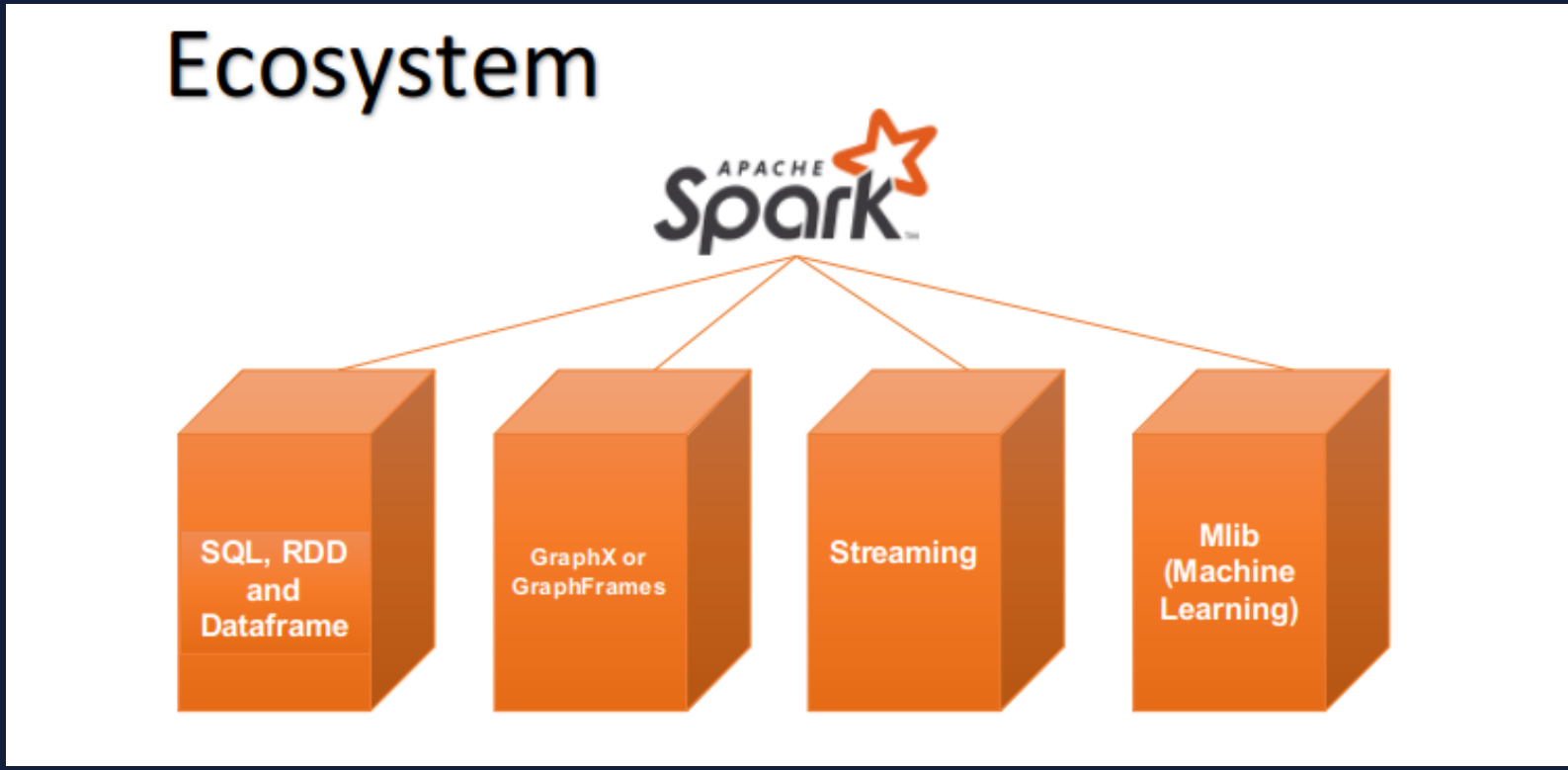
En esta lección aprenderemos a trabajar con Spark usando una librería de Python como es PySpark, para ello usaremos una imagen de Docker y veremos los distintos usos con los que podemos trabajar.

OBJETIVOS

Al finalizar esta lección serás capaz de:

1. Conocer que es Spark y sus usos en Big Data.
2. Usar la imagen de pyspark-notebook para trabajar con este entorno.
3. Conocer los usos de Pyspark librería de Python en Big data.

Spark


Ecosystem
APACHE
Spark
SQL, RDD and Dataframe
GraphX or GraphFrames
Streaming
Mlib (Machine Learning)

Spark SQL y DataFrame

Como trabajar con Spark gestionando los datos como DataFrame o como datos estrcuturados (SQL)

```
In [10]: #show first 5 rows
df.show(5)
```

```
+--------------+--------------------+------+------------+--------------------+-----------+-----+----------------
--+----------+---------+--------------------+------+--------------------+-----+
|Quarter Ending|          Department|UnitNo|Vendor Number|              Vendor|       City|State| DeptID Descripti
on|    DeptID|   Amount|             Account|AcctNo|    Fund Description| Fund|
+--------------+--------------------+------+------------+--------------------+-----------+-----+----------------
--+----------+---------+--------------------+------+--------------------+-----+
|    12/31/2019|Vt Housing & Cons...|  9150|   0000002188|Vermont Housing &...| Montpelier|   VT|              Tru
st|9150120000|1075000.0|Transfer Out - Co...|720010|Housing & Conserv...|90610|
|    12/31/2019|Vt Housing & Cons...|  9150|   0000375660|Wagner Developmen...|Brattleboro|   VT|            VT RE
DI|9150293000|   4612.5|Other Direct Gran...|552990|Housing & Conserv...|90610|
|    12/31/2019|Vt Housing & Cons...|  9150|   0000043371|Vermont Land Trus...| Montpelier|   VT|              Tru
st|9150120000|112916.67|Other Direct Gran...|552990|Housing & Conserv...|90610|
|    12/31/2019|Vt Housing & Cons...|  9150|   0000042844|University of Ver...| Burlington|   VT|Farm Viability-VH
CB|9150255000| 17152.74|Other Direct Gran...|552990|Housing & Conserv...|90610|
|    12/31/2019|Vt Housing & Cons...|  9150|   0000160536|Lahar Stephanie &...|
CB|9150255000|   4850.0|Other Direct Gran...|552990|Housing & Conserv...|90610|
+--------------+--------------------+------+------------+--------------------+-----------+-----+----------------
--+----------+---------+--------------------+------+--------------------+-----+
only showing top 5 rows
```

```
In [15]: spark.sql(
'''
SELECT `Quarter Ending`, Department, Amount, State FROM VermontVendor
WHERE Department = 'Education'
LIMIT 10
'''
).show()
```

```
[Stage 19:==============>                                          (1 + 3) / 4]
+--------------+----------+--------+-----+
|Quarter Ending|Department|  Amount|State|
+--------------+----------+--------+-----+
|    12/31/2012| Education|  302.12|   VT|
|    12/31/2012| Education|531548.0|   VT|
|    12/31/2012| Education| 14082.0|   VT|
|    12/31/2012| Education| 5337.66|   VT|
|    12/31/2012| Education|164436.0|   VT|
|    12/31/2012| Education|  8295.0|   VT|
|    12/31/2012| Education|   646.5|   VT|
|    12/31/2012| Education|    29.9|   VT|
|    12/31/2012| Education| 34159.0|   VT|
|    12/31/2012| Education|  2626.0|   VT|
+--------------+----------+--------+-----+
```

Spark RDD (Resilient Distributed Dataset)

Es una colección de objetos similar a la lista en Python

RDD

```
In [20]: sc = spark.sparkContext

In [22]: rdd = sc.parallelize(range(1000))
         rdd.takeSample(False, 5)

Out[22]: [629, 851, 269, 187, 190]

In [ ]:
```

CONCLUSIONES

1. Spark sirve para procesamiento de datos distribuidos diseñado para ser rápido

2. Spark SQL sirve para el procesamiento de datos estructurados ySpark DataFrame es una colección distribuida de datos organizados en columnas con nombre.

3. Spark RDD es una colección distribuida inmutable de objetos.

MUCHAS GRACIAS POR SU ATENCIÓN

eip ESCUELA INTERNACIONAL POSGRADOS

imaniega@grupomainjobs.com

Isabel Maniega
https://www.linkedin.com/in/isabel-maniega-cuadrado-40a8356b/

twitter.com/eiposgrados

facebook.com/eiposgrados

instagram.com/eiposgrados